Descripción del problema predictivo a analizar, análisis del DataSet, métricas y criterios.

Juan Diego Arroyave, Sebastián Sánchez

Problemática

Los accidentes de tráfico son un problema importante en el Reino Unido, con un promedio
de 5,5 millones de accidentes cada año. El costo social de los accidentes de tráfico es significativo, alcanzando los 23.000 millones de libras esterlinas al año.
El objetivo de este estudio es desarrollar un modelo predictivo que estime el número
de accidentes de tráfico en el Reino Unido en 2023. El modelo se basará en datos históricos enlistados en un Data Set de accidentalidad entre 2005 y 2012 para Reino Unido, hallado en la plataforma de Kaggle.

Importación y tratamiento de datos

Se importan los datos de accidentes de tránsito en el Reino Unido, entre los años 2005 a 2012, de una competición de Kaggle alojada en el siguiente url: https://www.kaggle.com/datasets/devansodariya/road-accident-united-kingdom-uk-dataset/data

De allí, se obtuvo alrededor de 33 millones de datos. Una vez se importa la base de datos, se comienza con el tratamiento del marco de datos, se identifican y eliminan datos duplicados, elimina columnas innecesarias (como la del índice de la base de datos) para poder realizar una correcta coerción de las variables.

En un principio, se evidenció que había una proporción del 0,23% de datos faltantes, pero dicho porcentaje no era real, ya que los datos registrado como ‘None’ los fueron tomados como entradas tipo (str), se decide reemplazarlos por datos vacíos obteniendo una proporción real de datos faltantes del 7,6%. Tras seguir analizando la naturaleza de las variables se aplicó una estrategia de imputación de datos faltantes: En la variable de condiciones especiales en el sitio y peligros en la calzada los datos vacíos fueron reemplazados por 'No special' y 'No hazards' respectivamente. Esto indicaría que no hay condición especial en el sitio ni peligros en la calzada. Además, los datos vacíos de la variable Time, fueron reemplazados por la hora 00:00 dejando así, una proporción de 1,48% de datos faltantes.

Estandarización de datos

Tras haber realizado la limpieza de los datos, dada la diferencia de escala de las variables se decide realizar un proceso de estandarización de algunas variables de interés del marco de datos:

- Si la policía atendió la escena del accidente
- Severidad del accidente
- Numero de vehiculos
- Número de victimas
- Límite de velocidad
- Área rural o urbana
- Día de la semana
- Hora
- Año

Lo anterior con el fin de evaluar un modelo con datos estandarizados y sin estandarizar. Al realizar el primer modelo se decide partir la BD para utilizar el 80% de los datos para entrenamiento y el otro 20% para evaluación.

Se plantea un modelo de regresión, en la cual se plantea la variable Time como variable respuesta y las demás variables como variables regresoras. Como resultado, se obtuvo un error al elegir una variable categórica como variable respuesta. Se pretende continuar con la exploración de diferentes modelos y dar con uno que se compagine bien con el desarrollo de la problemática